Graph-Based Representation of Borrowing Chains in a Web Portal for Loanword Dictionaries

Peter Meyer
Institut für Deutsche Sprache, Mannheim
meyer@ids-mannheim.de

Abstract

This paper reports on an ongoing lexicographical project that investigates Polish loanwords from German that were further borrowed into the East Slavic languages Russian, Ukrainian, and Belorussian. The results will be published as three separate dictionaries in the *Lehnwortportal Deutsch*, a freely available web portal for loanword dictionaries having German as their common source language. On the database level, the portal models lexicographical data as a cross-resource directed acyclic graph of relations between individual words, including German 'metalemmata' as normalized representations of diasystemic variants of German etyma. Amongst other things, this technology makes it possible to use the web portal as an 'inverted loanword dictionary' to find loanwords in different languages borrowed from the same German etymon. The different possible pathways of German loanwords that went through Polish into the East Slavic languages can be represented directly as paths in the graph. A dedicated in-house dictionary editing software system assists lexicographers in producing and keeping track of these paths even in complex cases where, e.g, only a derivative of a German loanword in Polish has been borrowed into Russian. The paper concludes with some remarks on the particularities of the dictionary/portal access structure needed for presenting and searching borrowing chains.
Keywords: online dictionary; graph databases; loanwords

1 Introduction

1.1 The Lehnwortportal Deutsch

The *Lehnwortportal Deutsch* (lwp.ids-mannheim.de) is a freely accessible online lexical information system developed at the Institute for German Language (IDS) that has been designed to provide unified access to a large number of both existing and newly produced XML-based dictionaries of German loanwords in other languages.[1] The modular architecture of the portal allows for easy integration of new resources of possibly very heterogeneous structure; each portal dictionary may have its own XML schema, as long as the underlying lexicographical information of the different constituent parts of an

1 The web portal in its present form has been developed in a project funded by the Federal Government Commissioner for Culture and the Media upon a Decision of the German Bundestag.

entry are unambiguously and explicitly encoded and separated in the markup, analogous to what is called the 'lexical view' in the TEI.dictionaries module, cf. Burnard, Bauman 2007, section 9.5 (online at http://en.guidelines.tei-c.org/html/DI.html#DIMVLV [04/11/2014]).

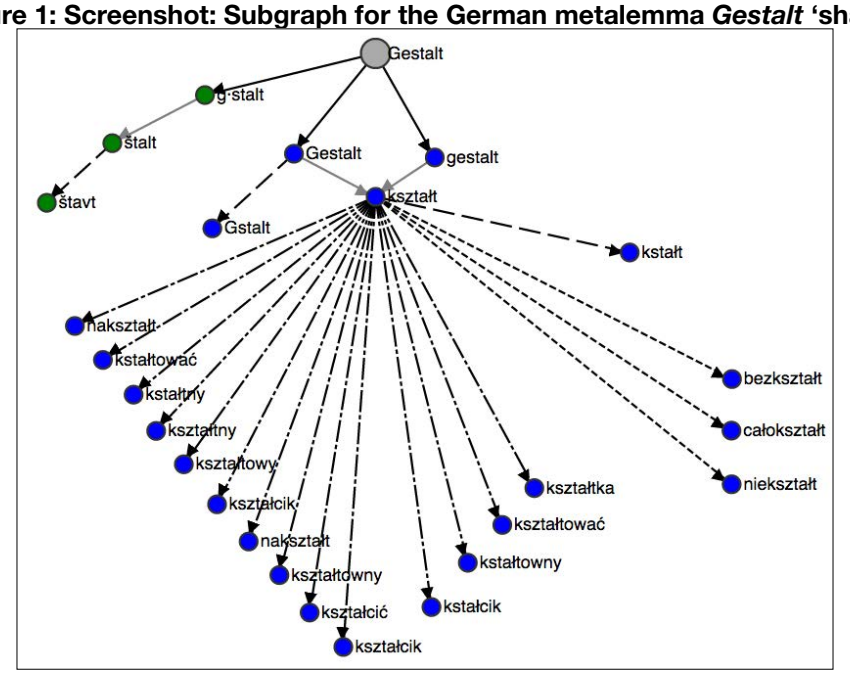
Apart from conventional access to the individual dictionaries, the portal offers complex cross-dictionary search functionality; in particular, it can be used as an 'inverted loanword dictionary' (Engelberg 2010) to trace the way of German words into different recipient languages, comparable to the manually compiled dictionary of Dutch loanwords in the world's languages by van der Sijs (2010). As any German etymon may appear in a variety of orthographical, phonetic/phonological and diasystemic variants in different entries within and across loanword dictionaries, these different forms are mapped in manual lexicographical work to etymologically corresponding 'normalized' word forms, wherever possible contemporary Standard German words. This is accomplished at the IDS with the help of an in-house software tool during the integration of a loanword dictionary into the web portal. These normalized entries, henceforth *metalemmata*, are used as headwords of the inverted loanword dictionary.

1.2 Graph-based Data Modeling

The XML-based representation of entries in the individual component dictionaries mainly serves as input for XSLT transformations that produce a fairly conventional, dictionary-specific HTML-based online presentation of the entries. For advanced search functionalities, however, a relational database is used that represents lexicographical information as a cross-resource network (a *directed acyclic graph*) of relations between words that are, as we say throughout this paper, 'recorded' in the individual dictionaries. These recorded words include metalemmata, etyma and loanwords alongside their variant forms, derivatives etc. Interactive visualizations of parts of this graph are available online; figure 1 below shows the subgraph for the German metalemma *Gestalt* 'shape'. Differently colored discs correspond to words recorded in different dictionaries (vertices/nodes in the graph); different kinds of relations (arcs/directed edges) are symbolized by different types of arrays between two discs. In the example we see that the 'normalized' contemporary German metalemma *Gestalt* 'corresponds to' [=dark solid arrow] a New High German etymon *Gestalt* and a Middle High German etymon *gestalt* as recorded in the portal's Polish loanword dictionary (color tag: dark blue) and to a Middle High German / Bavarian etymon *g·stalt* as recorded in the Slovene dictionary (color tag: green). We further see that, e.g., the etymon *gestalt* 'has been borrowed into' Polish [=grey solid arrow] as *kształt* which 'has a variant phonetic' form *kstalt* [=black long-dashed arrow] and from which (amongst other words) the verb *kształcić* 'has been derived' [=dashed-dotted arrow]. The relationships between words recorded in the same loanword dictionary entry are programmatically extracted from the XML source of the entry on a per-dictionary basis, making use of the fact that different kinds of relations correspond to different structural configurations in the entry document that depend on the XML schema of the dictionary and can be described using XPath expressions. Every word in the graph has a set of attributes (diasystemic, grammatical, semantic information) obtained by encoding the appropriate pieces of in-

formation as contained in the entries in a portal-wide unified data format. For more details, cf. Meyer (2013b; to appear).

Figure 1: Screenshot: Subgraph for the German metalemma *Gestalt* 'shape'



(http://lwp.ids-mannheim.de/art/meta/lemma/Gestalt).

1.3 Tracing the Way of Polish Loanwords from German into East Slavic

In a joint project funded by the German Research Foundation the Institute of Slavic Studies at the University of Oldenburg and the Institute of German Language (IDS, Mannheim) are currently developing dictionaries of German loanwords in the three East Slavic languages Russian, Ukrainian, and Belorussian that were mediated through Polish words recorded in the portal's dictionary on German loans in Standard Polish (previously published as a standalone resource: de Vincenz, Hentschel 2010). This endeavor draws on a rich Slavic tradition of historical lexicography; a wealth of (partially unpublished) dictionary material (starting with a basis of 15 historical and contemporary monolingual dictionaries of Russian, Ukrainian, and Belorussian) will be excerpted and analyzed both in Oldenburg and at the editorial offices of those dictionaries that are still work in progress, while the portal integration of the resulting dictionaries with an estimated total of 1900 new entries will be carried out in Mannheim.

The present paper focuses on data modeling issues and the technical and procedural specifics of dealing with (arbitrarily long) borrowing chains in the context of this project.

2 Modeling and Editing Borrowing Chains in the Portal's Graph Database

2.1 Data Modeling Aspects: Borrowing Chains as Paths in the Graph

Borrowing chains are the premier *raison d'être* for the graph-based data modeling in the *Lehnwortportal*. Figure 2 (below) shows, if only in a highly schematic fashion, the sample case of German *Drucker* 'printer' that has entered the East Slavic languages mostly through Polish. The different pathways obviously form a small directed graph that can be added more or less directly as a new subgraph to the portal data graph. The dashed arrows indicate less likely borrowing pathways; correspondingly, edges in the portal graph may be assigned weights to indicate likelihood of a borrowing relation and to calculate rankings of search results. Note that there are three different paths in this subgraph all leading from the German etymon *Drucker* to the Russian loanword *drukar'*.

Figure 2: Possible borrowing paths of German Drucker 'printer' into the East Slavic languages.

2.2 Some Technical Aspects of the Lexicographical Process

The graph data layer atop the XML resources of individual loanword dictionaries requires a highly specialized lexicographical process of its own (cf. Meyer, to appear). The graph is not a self-contained resource; instead, it must be constructed anew from the individual dictionaries and portal-specific cross-reference information as soon as any of the portal's data sources changes. Tracing borrowing chains complicates the picture considerably. For the project presented here, a complex in-house desktop dictionary editing software is being developed which allows lexicographers to collaboratively

compile and edit excerpts *using the lemmata (and other recorded words and meaning definitions) of the portal's Polish loanword dictionary* (de Vincenz, Hentschel 2010) *as a common frame of reference*. Figure 3 (below) shows a screenshot of a preliminary version of the editor used for excerpting. The working lexicographer selects a Polish loanword from (de Vincenz, Hentschel 2010) such as *browar* 'brewer; brewery' from Middle High German *brouwer* 'brewer' (1). A preview of this entry is displayed in the main window (2). All hitherto produced excerpts of existing dictionary entries on East Slavic borrowings from the selected Polish word are listed as a tree structure in the editor (3); for each such entry, the tree shows all recorded (phonetic and diasystemic) variants, meanings, derivatives (with their own range of variants and meanings) and competing near-synonyms that have been input so far. Clicking on a tree item (here, on the variant *provar* of the entry *brovar* in the multivolume Belorussian Historical Dictionary *Histaryčny sloŭnik belaruskaj movy*) opens an input panel (4) for all pertinent lexicographical information, including an arbitrary number of records and quotations. A preview of the current state of the whole excerpt is also available (5). There is a separate input panel, not shown in figure 3, for editing cross-dictionary information on the possibly multiple borrowing paths within the East Slavic languages. Often Polish loanwords from German have formed compounds and derivatives; it is well possible that only one of these derived forms, but not the 'original' loanword, has been passed on into an East Slavic language. The editing software also offers convenient input options for such cases. There are additional tools for compiling the entries of the three new East Slavic loanword dictionaries from the excerpts.

There are many reasons why an off-the-shelf software solution would not have been suitable for the lexicographical tasks of the project. To begin with, it would have been next to impossible to customize a commercial dictionary editing application in order to incorporate cross-referencing functionality to an existing dictionary. In this particular case, cross-references are needed not only to whole entries of the Polish dictionary (de Vincenz, Hentschel 2010), but also to derivatives and compounds recorded in these entries, and, most important, to the different word senses given in the entries since they will serve as a *tertium comparationis* for word sense distinctions in the East Slavic loanwords. It would have been possible to customize a professional XML editor by implementing some kind of cross-referencing plugin. However, there is another layer in the editing process that cannot easily be managed in XML: After compiling excerpts of entries on a German loanword in, say, Ukrainian, in a number of Ukrainian loanword dictionaries, these excerpts have to be merged in a rather complex way to produce a new entry in the Ukrainian loanword dictionary of the portal. The excerpted loanword dictionaries (which may or may not cover different periods of the language) will have differing lemmatizations, list different variants of the word, use incompatible word sense distinctions and so on. On the other hand, there will usually be a lot of duplicate information. As a consequence, the amalgamation process of creating entries in the three new East Slavic portal dictionaries is far from trivial; doing this by cutting XML fragments from the excerpts and pasting them into the XML structure of the newly created entries would be an excessively tedious, error-prone and confusing task, even more so since word sense distinctions in parallel entries in the three dictionaries should be made in as uniform a

manner as possible, based on the distinctions in the entries of (de Vincenz, Hentschel 2010). It is not a realistic goal to develop software tools for these tasks as simple XML editor plugins; even the very idea of using XML as the basic frame of reference is problematic in such a complex cross-resource editing context.

In fact, the software developed at the IDS is not directly XML-based, but uses a straightforward object-oriented data model for both the excerpts and the newly produced entries. This greatly simplifies the underlying cross-referencing and the implementation of tools for merging and validating lexicographical data from a large number of resources. The software produces XML serializations of the data that can be used both to construct HTML views of the data and to define the directed graph of the portal.

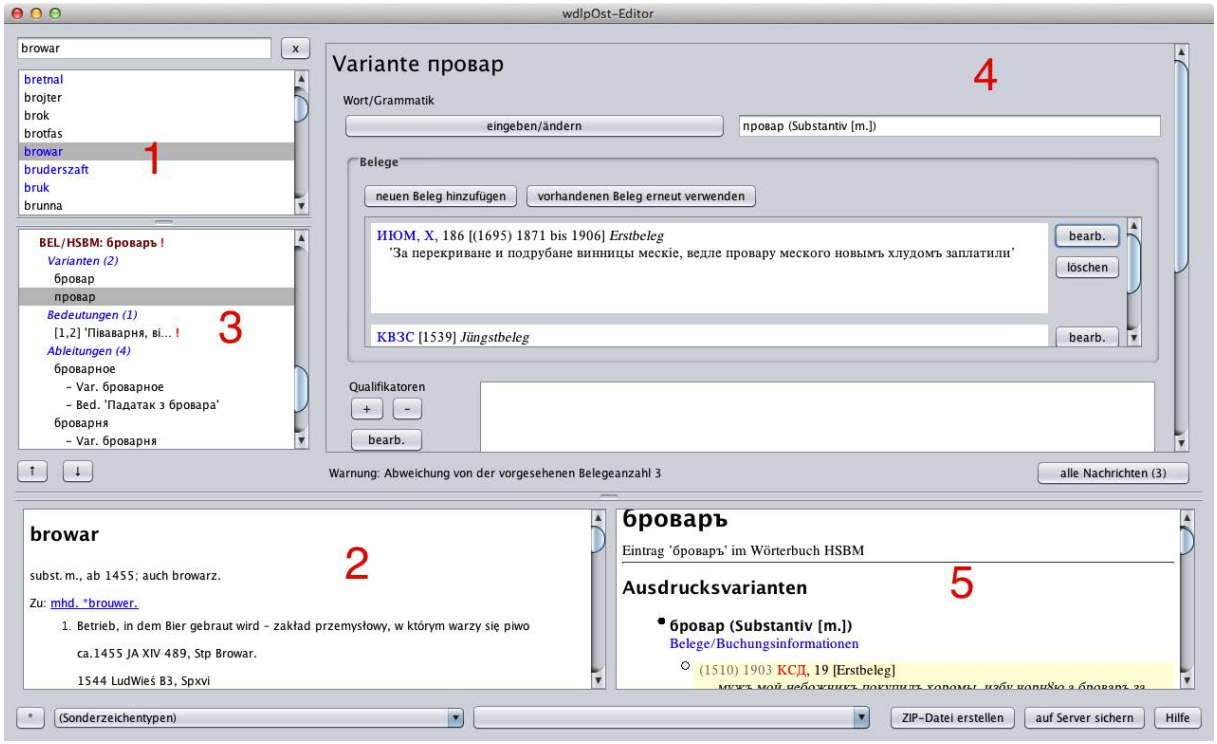


Figure 3: Screenshot: Preliminary user interface of the dictionary writing software.

2.3 General Lexicographical Issues Concerning Borrowing Chains

As explained above, the project presented in this paper strives to compile new loanword dictionaries that directly reference a Polish loanword dictionary already integrated into the portal. As the *Lehnwortportal* has been designed with a focus on leveraging existing resources, we expect to see other cases in the future where information from multiple existing loanword dictionaries is combined to reconstruct borrowing chains. Here, the data graph is an ideal means of abstracting from micro- and mediostructural idiosyncrasies of the dictionaries involved. Dutch may serve as a good example since

it has served as a 'hub' that mediated German lexis into many languages, in particular those of colonized countries. Thus, we might try to combine information on Dutch loans from German – as represented in a traditional loanword dictionary (e.g., van der Sijs 2005) – with information on Dutch loanwords in other languages – as given in (van der Sijs 2010) – to (re)construct borrowing chains from German via Dutch into other languages. In these cases, an 'intermediate' Dutch loan corresponds to *two* connected vertices in the graph as it appears in two independent lexicographical resources – both as a loanword from German and as an etymon for Dutch loans in other languages. The etymological identification of these two 'instances' of the intermediate loanword is part of the lexicographical process to be carried out for the portal. There are many technical and lexicographical issues arising in borderline cases, e.g., when borrowing chains are directly specified in a loanword dictionary entry (such as "from German *Drucker* via Polish *drukarz*"). Here, multiple cross-resource etymological identifications with words in other resources might become necessary, even with the possibility of conflicting information, e.g. if another loanword dictionary of the portal gives a different German etymology for an intermediate loan.

3 Access Structures for Borrowing Chains in the Portal

3.1 Online Entry Presentation

The information on borrowing chains is, in many cases, not present within the confines of a single loanword dictionary entry, but is instead distributed between different resources. The online presentation of individual dictionary entries should nevertheless make this information visible in all of the individual dictionary entries involved. At present, loanword dictionary entries and entries of the 'inverted loanword dictionary' of German metalemmata systematically cross-reference each other in the *Lehnwortportal*; in the case of 'indirect' loanwords from German, the web application will use the data graph to add another layer of information, viz. on the 'intermediate' or 'terminal' loanwords, to the existing entries in the Polish and East Slavic dictionaries. These additions only concern the presentation layer; the underlying entries remain unaltered. A special feature of the presently compiled East Slavic dictionaries will be the presence of cross-dictionary commentaries (including schematic visualizations of borrowing pathways) on all entries that refer to the same Polish loanword, since often German loanwords in an East Slavic language could also have been borrowed directly from German or were mediated by another East Slavic language (cf. figure 2 above).

3.2 Advanced Search Options

It is highly desirable that portal users can include search criteria concerning borrowing chains into their advanced queries such that, e.g., German loanwords in language X that were possibly mediated

through language Y can be found. The *Lehnwortportal* offers fairly advanced and granular search options (cf. Meyer 2013a) that allow the inclusion of complex criteria concerning both German etyma (including metalemmata) and loanwords. With the inclusion of the Slavic dictionaries, search criteria will also be attributable to 'intermediate' loans in a borrowing chain. Search results will be ranked according to the weight of the edges of the graph path. Borrowing paths will be specifiable through a planned extension of the declarative domain-specific query language that is currently available for advanced portal users (figure 4); cf. (Wood 2012) for a general overview on graph database query languages and (Meyer 2013a) for more information on the portal's query language. Even a graphical search through an interactive visual query language for graph databases is conceivable (cf. Blau et al. 2002) and would allow users to literally draw the borrowing paths they are looking for.

Figure 4: Screenshot: Example search using the portal's graph query language (http://lwp.ids-mannheim.de/search/prof).

For illustration purposes, here is a rough preview of how a simple query involving borrowing chains will look like in the portal's query language. Note that the original query language has a German-like context-free grammar; here, we present a corresponding English-like version. The query reads: "Find all Ukrainian or Belorussian words (including variants, derivatives etc.) in the database that have

been borrowed through Polish from a German noun ending in *ung*." The query uses graph-theoretical terms where appropriate; thus, any loanword borrowed from German is represented by a node in the portal's directed graph that is a *descendant* of the node corresponding to the German etymon. The results of the query are ordered triples (germanWord, polishWord, eastSlavicWord) of those words recorded in entries of the portal's dictionaries that comply with all of the constraints specified in the query.

find etymon germanWord.
find loanword polishWord.
find loanword eastSlavicWord.
the language of germanWord is German.
the language of polishWord is Polish.
(the language of eastSlavicWord is Ukrainian OR the language of eastSlavicWord is Belorussian).

germanWord is a noun.
germanWord ends in 'ung'.
polishWord is descendant of germanWord.
eastSlavicWord is descendant of polishWord.

4 References

Blau, H., Immerman, N. & Jensen, D. (2002). *A Visual Language for Querying and Updating Graphs*. Technical Report 2002-037. University of Massachusetts, Amherst.

Burnard, L., Bauman, S. (eds.) (2007). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Charlottesville, Virginia: TEI Consortium. Online: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html [04/11/2014].

de Vincenz, A., Hentschel, G. (2010). Wörterbuch der deutschen Lehnwörter in der polnischen Schrift- und Standardsprache. Von den Anfängen des polnischen Schrifttums bis in die Mitte des 20. Jahrhunderts. (= Studia slavica Oldenburgensia, vol. 20). Oldenburg: BIS-Verlag. Online edition: http://diglib.bis.uni-oldenburg.de/bis-verlag/wdlp [04/11/2014].

Engelberg, S. (2010). An inverted loanword dictionary of German loanwords in the languages of the South Pacific. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress (Leeuwarden, 6-10 July 2010)*. Ljouwert (Leeuwarden): Fryske Akademy, pp. 639-647.

Meyer, P. (to appear). Von XML zum DAG: Der lexikographische Prozess bei der Erstellung eines graphenbasierten Wörterbuchportals. In F. Mollica, M. Nied, M.J. Domínquez Vazquez (eds.) *Zweisprachige Lexikographie im Spannungsfeld zwischen Translation und Didaktik*. (to appear in: Lexicographica, Series Maior).

Meyer, P. (2013a). Advanced graph-based searches in an Internet dictionary portal. In I. Kosem, J. Kallas, P. Gantar, P. Krek, M. Langemets, M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 488-502. Accessed at: http://eki.ee/elex2013/proceedings/eLex2013_34_Meyer.pdf [04/11/2014].

Meyer, P. (2013b). Ein Internetportal für deutsche Lehnwörter in slavischen Sprachen. Zugriffsstrukturen und Datenrepräsentation. In S. Kempgen, N. Franz, M. Jakiša, M. Wingender (eds.) *Deutsche Beiträge zum 15. Internationalen Slavistenkongress, Minsk 2013*. München: Otto Sagner, pp. 233-242. (=Die Welt der Slaven. Sammelbände, vol. 50).

Meyer, P., Engelberg, S. (2011). Ein umgekehrtes Lehnwörterbuch als Internetportal und elektronische Ressource: Lexikographische und technische Grundlagen. In H. Hedeland, Th. Schmidt, K. Wörner (eds.) *Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011*. Hamburg: Universität Hamburg, pp. 169-174 (=Arbeiten zur Mehrsprachigkeit/Working Papers in Multilingualism, Series B, No. 96).

van der Sijs, N. (2005). *Groot leenwoordenboek*. Utrecht: Van Dale Lexicografie.

van der Sijs, N. (2010). *Nederlandse woorden wereldwijd*. Den Haag: SDU Uitgever.

Wood, P. T. (2012). Query Languages for Graph Databases. In *SIGMOD RECORD*, 41(1) (March), pp. 50-60.